# Project 3: Relation Extraction

Alexander Sifel - Ana Terović - Lukas Hofstetter

TUW - Natural Language Processing — January 26, 2023

## Data-Sets

Instances 609
Features 4
Selected Features 3
Classes 2

Food disease

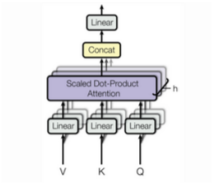
7670 Instances
18 Features
4 Features selected
2 Classes

Crowed truth

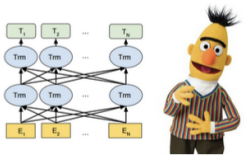
## Algorithms

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Naïve Bayes (NB)




Multi Head (MH)



Bert with head (B)

POTATO

Potato (P)

## Results

| | Naive Bayes | | finetuned Bert | | Potato | | | |
|---|---|---|---|---|---|---|---|---|
| | micro avg | macro avg | micro avg | macro avg | is_cause | | is_treat | |
| accuracy | | | | | 0.77 | 0.5932 | 0.5932 | 0.5593 |
| precision | 0.81 | 0.90 | 0.89 | 0.88 | 0.8571 | 0.6154 | 0.6154 | 0.2272 |
| recall | 0.64 | 0.48 | 0.85 | 0.83 | 0.7272 | 0.7272 | 0.7272 | 0.3571 |
| F1 | 0.71 | 0.49 | 0.87 | 0.85 | 0.7869 | 0.6667 | 0.6667 | 0.2778 |
| support | 47 | 47 | | | | | | |
| | | | | | trainer1 | trainer2 | trainer1 | trainer2 |

Table 1: algorithm comparison table

# Examples and Conclusions

## Naive Bayes

In our first attempt we implemented a Naive Bayes classifier using the *sklearn* library as our baseline, achieving an *F1-score* of 71% in *micro_avg* and 49% in *macro_avg*.

**is_cause predictors**
Positive: ['atopic', 'sensitization', 'well-known', 'exposed', 'rhinitis', 'allergen', 'frequently', 'inhalation', 'baker', 'worker']
Negative: ['treatment', 'compound', 'protective', 'ad', 'component', 'model', 'medicine', 'traditional', 's', 'various']

**is_treat predictors**
Positive: ['bioactive', 'effective', 'therapeutic', 'chinese' 'benefit', 'anti-inflammatory', 'anti-disease_entity', 'reduces', 'medicine', 'traditional']
Negative: ['virus', 'worker', 'caused', 'ibv', 'industry', 'case', 'outcome', 'cause', 'dust', 'production']

The main problem referring to the metrics is that the recall for the **is_cause** model was super low (7 %). The data-set is imbalanced with less positive examples for the **is_cause** relation, so it might be because of a bias. One potential mitigation for this could be using stratification or re-weighting.

## Multi-head attention

For Milestone 2 we have implemented Multi-head attention model. The motivation for implementation was the fact that the multi-head model gives the transformer greater power to encode multiple relationships and nuances for each word. The model was tested on the *CrowdTruth* data-set and it gave very poor results with accuracy just above 50%. We will later on realize when looking into the data-set that the poor results weren't due to this models ability but because of the quality of the data-set, therefore we excluded it in further investigations.

## Bert fine-tuned

For the implementation of the final project we decided to start with a transfer learning approach using Googles BERT model (checkpoint: emilyalsentzer/Bio_ClinicalBERT) with a classification head of 2 linear layers, achieving an F1-score of 87% in micro_avg and 85% in macro_avg.

**is_cause false positives**
*Since influence has been related to the development of chronic condition prevalent in the western world, the use of sweeteners has gradually increased worldwide over the last few years.* - misslabeled

**is_treat false positives**
*however, the validity of influence as a treatment for condition (ra), an autoimmune disorder, has not been confirmed yet*" - confusion by counterfactual phrasing

## Potato

For our white-box model we have implemented knowledge graphs and rule sets using the *POTATO* framework. For the first model we have allowed the rules to consist of just the terms resulting in much more generalized rule set. Also, the rules the model suggested were quite explainable. For is_cause relation, signaling words were "increase", "patient", "symptom" while for is_treat relation they were "reduce", "decrease", "against", "prevent", "improve". This has given us an F1-score of 79% for the is_cause relation and 67% for the is_treat relation. These were quite good results when we compare them to a black-box model BERT. For the second model we have made restrictions on the number of edges, not allowing the model to suggest rules consisting of just one term. This resulted in more specific rules which as expected gave poorer results. Is_cause relation had an F1-score of 67% and is_treat relation a score of 28%. All together, with POTATO we were able to achieve good results and also understand exactly why the models made the decisions that they made.

## Contributions

Ordered contributions per magnitude (from left to right)

- Milestone 1: Alex
- Milestone 2: Ana, Lukas
- Milestone 3: Alex/Ana, Lukas
- Presentation: Alex, Ana, Lukas
- Management Summary: Lukas, Ana, Alex